修士論文の和文要旨

大学院情報システム学研究科　博士前期課程　情報システム設計学　専攻		
氏　　名	木村　洋章	学籍番号　0450011
論　文　題　目	セグメントの圧縮性を用いた文書の自動要約と分類	

要　旨

ITの発展により誰もが情報を発信･受信できるようになった．このように情報が増えつつある状況の中，膨大な文書データを処理する技術への関心が高まっている．たとえば検索システムを利用することで，瞬時にして特定キーワードを含む文章を世界中のWEBページから検索することが可能となった．しかし単純なキーワード検索では検索結果を絞り込むためにキーワードを増やしたり，表記の揺らぎによる検索漏れを防ぐためにキーワードを変えたりする必要があり，検索の意図を伝えるのが難しい．

そこでキーワードに代わって関連する文章を入力することで類似した文章を検索する文書連想検索技術が近年活発に研究されている．この技術によりキーワード検索では伝えきれなかった検索の意図を伝えることができ，話題の絞り込みを行えると期待できる．また検索結果を話題ごとに分類し，それらの要約を提示することで全体を読まずに検索結果を絞り込め，素早い検索が可能となる．また流行な話題を検索者に伝えるシステムも注目されている．

しかし，このような高レベルな情報処理技術には人手によるシステムのメンテナンスや，多言語への対応，言語･文法知識の導入が不可欠であり，情報変化への対応ができない．なぜなら情報は時間とともに変化し，事前知識が役に立たなくなることがあるからである．特に情報の新規性は日々変化していくため，流行を見極めるのは難しい．

そこで我々は，膨大かつ時間とともに変化する情報を対象とするシステムには，人手を要せず，下記の機能が必要と考える．

(1) 統一かつ簡便な枠組みで分類や認識を実現できること
(2) 新規／公知情報の自動検出を実現できること
(3) 文書の要約が実現できること

本論文では，圧縮性に基づいて情報を特徴づけるパラダイムであるPRDCを足場にこれらの問題を解決する可能性を探った．その結果，(1)については，PRDCという単一原理で文書分類や類似性判定を，言語解析知識などを用いずに行う手法を提案した．実験では迷惑メール分類タスクに適用して現在主流であるベイジアンフィルタよりも高精度に迷惑メール分類を行えることがわかった．(2)については，既存文書から構成される基底辞書空間上で分類できない情報を新規であると自動判定できることを示した．(3)については(2)を応用して新規性･公知内容を持つ断片文章を抽出することで自動要約法を提案し，人間による主観評価を行い有効性が確かめられた．